Concatenative Speech Synthesis: A Review

Rubeena A. Khan
Research Scholar
J.J.T.U
Jhunjhunu, India

J. S. Chitode, PhD
Professor: Dept. of Electronics
V.I.T
Pune, India

ABSTRACT

The primary objective of this paper is to provide an overview of existing Concatenative Text-To-Speech synthesis techniques. Concatenative speech synthesis can be broadly categorized into three categories, Diphone Based, Corpus based and Hybrid. Diphone based speech synthesis relies on different signal processing techniques such as PSOLA, FD-PSOLA etc. These signal processing techniques introduce unwanted artifacts in the synthesized speech. The most popularly used method is the Unit selection synthesis which is a corpus based synthesis method. This method produces the most natural sounding synthetic speech.

General Terms

Speech processing

Keywords

TTS; PSOLA; TD-PSOLA; FD-PSOLA; ESNOLA; MOS; SUS; DRT; HMM.

1. INTRODUCTION

Speech Synthesis is a technique that converts text into machine generated speech waveforms [1]. There are basically three methods by which TTS systems can be built: Articulatory, Formant and Concatenative synthesis. In Articulatory synthesis speech is generated by trying to model the human articulators like the lips, tongue, velum, pharynx, vocal cord, nose etc. Typically 7-11 parameters for sufficiently describing the different articulatory motions. [2] Speech synthesized by this method requires huge computational costs hence it is not practically used. Formant synthesis is based on the source-filter model of speech production. An excitation produced by a source, passes through a filter that models the vocal tract. The excitation source can either be an impulse train representing voiced sound or random noise representing unvoiced sound. The excitation source is modified by the resonances of the vocal tract in order to produce speech. Concatenative synthesis produces artificial speech by concatenating prerecorded units of speech such as phonemes, di-phones, syllables, words or sentences [3]. In order to achieve intelligent and natural sounding speech, multiple instances of speech units are stored in an inventory. Hence this method is also known as corpus based speech synthesis.

The remainder of the paper is organized as follows. The first section gives an overview of Concatenative speech synthesis. Followed by an overview and comparison of different types of Concatenative TTS systems. The next section discusses available databases for speech synthesis systems. The seventh section describes the different performance parameters of TTS systems. This is followed by the depiction of Speech Synthesis Frameworks and finally the conclusion.

2. CONCATENATIVE SPEECH SYNTHESIS

Basically Concatenative speech synthesis systems involve the selection of appropriate units from the speech inventory, algorithms that join the selected units and some signal processing so as to smoothen the concatenation boundaries [4]. In concatenative synthesis speech is produced by selecting and concatenating appropriate speech units from the speech database. The speech database can have speech units of different sizes such as phones, di-phones, syllables, words or sentences. The size of the speech units stored affects the quality of synthesized speech, if large sentences are stored the speech synthesized will sound natural, but it will restrict the flexibility of the TTS. Whereas if small units such as phones are selected it will provide more flexibility but with degraded quality [5]. Hence selection of an appropriate unit is very essential. The major factors influencing the quality of synthesized speech are the continuity of acoustic characteristics at the concatenation point, such as fundamental frequency, amplitude, speaking rate and the availability of speech units having appropriate prosody in the database [5].

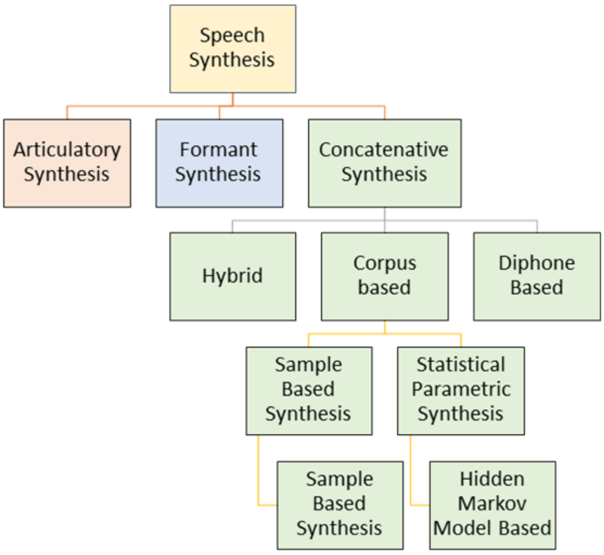

Fig 1: Taxonomy of Speech Synthesis Methods

3. DI-PHONE BASED SPEECH SYNTHESIS

This method of speech synthesis uses di-phones as the basic speech unit [6]. A di-phone consists of two connected half phones starting in the middle of first phone and ending in the middle of second phone. In di-phone based synthesis only single instances of all speech units are available in the speech inventory [5], hence to obtain good quality synthesized speech with the desired prosody, various signal processing methods are applied [3]. Some of the signal processing methods such as PSOLA, TD-PSOLA, and LP- PSOLA etc are summarized in the Fig 2.

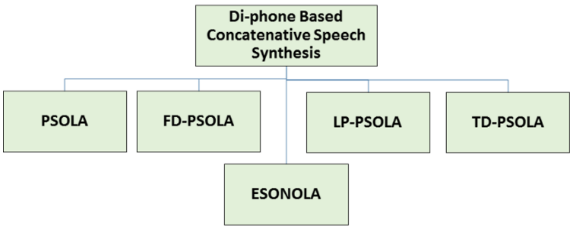

Fig 2: Various Signal Processing Techniques Used in Di-phone Based Speech Synthesis

3.1 Pitch Synchronous Overlap-Add (PSOLA)

It is an analysis-synthesis method where the speech signal is decomposed into a number of pitch-synchronous short-term waveforms. These short term signals can be altered either in the time domain or spectral domain to obtain multiple synthetic short-term signals. Synthetic speech is produced by overlapped addition of these short-term signals [5]. PSOLA does not lose any information of the signal since it works directly on the signal without using any model. It has the advantage of preserving the spectral envelope when pitch shifting is used [19].

3.2 B. Frequency Domain Pitch Synchronous Overlap-Add (FD-PSOLA)

It is used for modifying the spectral envelope; it uses linear prediction to compute the spectral envelope. The pitch is modified by linear interpolation of the spectral envelope. The main advantage of this method is the ease of implementation since the operations are performed in the frequency domain. Since the phase continuity cannot be controlled explicitly in FD-PSOLA it leads to discontinuities at the concatenation boundaries when used in concatenative speech synthesis [22]

3.3 Linear prediction pitch synchronous overlap and add (LP-PSOLA)

In this method the modification of pitch and duration of the speech signal is achieved by manipulating the LP residual [23]. It is suitable for pitch-scale modification, since it provides independent control over the spectral envelope for synthesis. The LP-PSOLA method has the disadvantage of producing phase mismatches and audible distortions due to overlap and adding of the windowed residual segments.

3.4 Time Domain Pitch Synchronized Overlap Add (TD-PSOLA)

It can be used in di-phone based synthesis for the modification of the prosody of speech waveforms, as it facilitates high quality pitch and time scale modifications [18] [17]. The TD-Psola has the advantage of being computationally efficient, but with the drawback of requiring a large speech data-base [17] and the quality of synthesized speech is affected by the detection of epochs in the speech signal which is very difficult to achieve in real time applications.

3.5 Epoch Synchronous Non Overlap Add (ESNOLA)

In this method the synthesized speech signal is generated by concatenating basic speech segments at the epoch positions of the voiced speech, where epochs represent vocalic or quasi-periodic sounds. ESNOLA can be used for pitch and duration modification of the synthesized speech [16] ESNola allows selection of smaller parts of a phoneme called partnemes as concatenation units , hence this reduces the size of the speech inventory[20]. ESONOLA supports the introduction of jitter shimmer and complexity perturbations which leads to naturalness in phonetic quality of synthesized speech [21].

3.6 Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

4. CORPUS BASED SPEECH SYNTHESIS

This technique of speech synthesis uses a data driven approach [3]. It depends on the availability of a good speech inventory, having good phonetic and prosodic features for the language under consideration [28]. The major issues related to corpus based approach is the segmentation and labeling of the speech inventory, which can be achieved by using automatic segmentation algorithms [29]

4.1 Unit Selection Synthesis

In this approach multiple instances of speech units having different prosodic features are stored. This approach is known as Unit selection based Concatenative TTS. An appropriate unit is selected from the database based on namely two costs. A target cost and a concatenation cost.

Target Cost: It estimates how similar the features of a database speech unit are to the features of the desired speech unit (candidate unit) [7]. The target cost comprises of target sub-costs. Each target sub cost is a cost of a single attribute of a speech unit such as energy, pitch etc[1]. The target cost can be calculated as:

$$Ct\,(ti,vi) = \sum pj=1\ \ wtj\ Ctj(ti,vi) \qquad (1)$$

Where ti is the target unit Vi is the candidate unit, p is the number of sub-costs used. Ctj is the jth target sub-cost, Wtj it is the weight given to the jth target sub-cost [8].

Concatenation Cost: It is a measure of how well two speech units join and match each other when they are concatenated [7]. The concatenation cost also comprises of multiple sub-costs. Each of these sub-costs is related to a specific continuity metric such as spectral continuity etc. The concatenation cost can be calculated as:

$$Cc\,(vi\text{-}1,vi) = \sum qj=1\ \ wcj\ Ccj(vi\text{-}1,vi) \qquad (2)$$

Where vi-1 and vi are candidate speech units for the (i-1)th and ith target speech units, q is the total number of subcosts used and Wcj is the weight associated with the subcost Ccj [9].

An exhaustive search is performed so as to select optimum speech units from the speech database. The Viterbi search is frequently used [11], to select the units to be concatenated from the speech inventory, so as to reduce both the target cost and the concatenation costs.

4.2 Statistical Parametric Synthesis

Statistical parametric synthesis makes use of averaged acoustic inventories that are extracted from the speech corpus [14]. The most commonly extracted parameters of speech are the spectral parameters such as cepstral coefficients or line spectral pairs, and excitation parameters, such as fundamental frequency [24][26] . Statistical Parametric synthesis has the advantages of requiring less memory to store the parameters of the model, rather than the data itself and it allows more

Variation in the speech produced for example, an original voice can be converted into another voice [3]. The most commonly used statistical parametric speech synthesis technique is the Hidden Markov Model (HMM) synthesis.

Hidden Markov Model (HMM) synthesis.

HMM synthesis consists mainly of two phases a training phase and a synthesis phase. During the training phase speech parameters are extracted from utterances in the speech training database and they are modeled as HMMs. In the synthesis phase the words to be synthesized their corresponding HMMs are identified from the database and parameters are extracted from these HMMS. Finally speech is synthesized from these extracted parameters. HMM based parametric speech synthesis is flexible since speech is stored in the form of parameters and it is easy to modify these parameters, but it has the disadvantage of poor quality in the naturalness of the synthesized speech. This is due to over smoothing of the parameters in the statistical model [25].

5. HYBRID TEXT TO SPEECH SYNTHESIS

The Hybrid TTS approach is a combination of the two main approaches of synthesis namely Concatenative synthesis and Statistical Synthesis. The hybrid TTS combines the characteristic of smooth transitions between adjacent speech segments of a Statistical TTS with the naturalness of a Concatenative TTS. This is achieved by interweaving natural speech segments and statistically generated speech segments. The statistical segments are positioned so as to smooth discontinuities in the synthesized speech, while enabling as far as possible natural speech sequences as they appear in the training inventory [15].

Table 1. A comparison of concatenative speech synthesis methods

METHODS	Sub-methods	Features
DI-PHONE BASED	PSOLA	• Preserves the spectral envelope when pitch shifting is used [19]
	FD-PSOLA	• Frequency Domain • Ease of Implementation • Can perform spectral modification. • Computationally intensive and has high memory requirements for storage • Discontinuities at the concatenation boundaries [22]
	LP-PSOLA	• Suitable for pitch-scale modification • Produces phase mismatches and audible distortions.
	TD-PSOLA	• Time Domain • Facilitates high quality pitch and time scale modifications. • Drawback of requiring a large speech data-base[17]
	ESONOLA	• Can be used for pitch and duration modification
CORPUS BASED	Unit Selection Based	• Is sample based. • Requires minimum or no signal processing. • Good quality in naturalness of synthesized speech [11],[27]. • Requires a large speech database.
	Hidden Markov Model Based	• Parameter based. • Muffled synthesized Speech [27]. • Unnatural speech. • Small speech inventory [15]. • More Signal processing required.
HYBRID		• Combination of Concatenative and Statistical TTS. • Degradation in speech quality when CTTS speech inventory is small [15]. • Requires signal processing

6. AVAILABLE DATABASES FOR SPEECH SYNTHESIS SYSTEMS

Some of the available speech corpus databases in different languages are summarized below.

LDCIL: Is a data consortium for 22 Indian languages such as Hindi, Marathi, Gujarati, and Konkani etc. The consortium has around 3112 hours of raw data and 15,362 hours of segmented data by multiple speakers, collectively for all the 22 Indian languages [30].

CMU ARTIC: It was designed keeping in mind the unit selection synthesis as target technology [31][32]. The database consist of around 1150 utterances .It includes US English male and female speakers.

CMU_FAF: This database consists of Facts and Fables It is recorded by a single US male speaker and contains 107 paragraphs .The database is automatically labeled [33].

CMU_SIN: is a database of speech in noise designed for use in unit selection speech synthesis research. It has 500 utterance subset from the CMU Arctic database. The database is recorded by a single male US English speaker. There are two versions of these recordings: one where the speaking style is speech in noise (sin) and one with normal speaking style (swn) [34].

KED TIMIT: This database contains 453 utterances spoken by a US male speaker. It is useful general database for simple prosody modeling and unit selection [35].

KAL DIPHONE: This database consists of a set of nonsense words containing all phone-phone transitions for US English.

ELRA-S0342: It is a database for Polish concatenative speech synthesis .This database consists of 1443 nonsense words including all the di-phones for the Polish language. Database is recorded by a female speaker. A 16 kHz sampling frequency and 16 bit resolution was used. The total duration of the recordings is 1.27 hours.

7. PERFORMANCE OF TTS SYSTEMS

The performance of speech synthesis systems involves the evaluation of the quality of synthesized speech, for naturalness and intelligence. This evaluation can be done on the basis of different subjective and objective scores.

7.1 Subjective Tests

Subjective tests can be used for testing how effective a synthetic voice is in a product [10]. They measure the listener's opinion of the synthetic speech.

a) Mean Opinion Score (MOS): It is the most popular test to measures the listener's opinion of the synthetic speech. The test involves a large number of participants. They listen to a set of synthesized sentences and rate them on a 5 point scale (excellent - bad) [11],[1].

b) Degradation of Mean Opinion Score (DMOS): This test can be used for testing the speaker similarity. In the DMOS test, subjects are made to listen to the target speaker's natural speech and a test sample (same sentence), they were asked to rate it a similarity score from the five-point Likert scale (5: exactly the same – 1: completely different) [38].

7.2 Objective Tests

Objective tests measure how the synthetic voice differs from the natural utterance. They can be used as diagnostic tools for refining the synthetic speech [10].

a) Semantically Unpredictable sentences (SUS) test: This test is used to evaluate the sentence level intelligibility of TTS synthesis, based on semantically unpredictable sentences. It uses sentences that are syntactically acceptable but semantically anomalous [37], for example "He ate the car".

b) Diagnostic Rhyme Test (DRT): This test is used to evaluate the intelligibility of TTS synthesis [36][1]. The test uses monosyllabic words having a consonant-vowel-consonant pattern and it measures the capability of discrimination of the initial consonants for the system under consideration [1] . Listeners listen to monosyllabic words which differ only in the first consonant and have to choose which word they have heard from pairs (for example, pin/fin, hit/bit).[11]

c) Mel-Cepstral Distortion (MCD): MCD is calculated between the original and synthesized speech. A low MCD indicates good quality of synthetic speech [12].MCD compares the Mel Cepstral Coefficients of two waveforms for finding the similarity between them. The mel-cepstral distortion (MCD), is defined as an extension of the simple Euclidean norm [13]. MCD can be used to estimate the quality of voices in new languages.

8. SPEECH SYNTHESIS FRAMEWORKS

8.1 MBROLA

MBROLA: it is a free for non-commercial use synthesis engine. The speech synthesizer is based on concatenation of di-phones. Input to this synthesizer is a list of phonemes along with prosodic information such as duration and pitch of phonemes and the output is speech samples. The MBROLA engine does not accept raw text as input hence it is not a Text-To-Speech synthesizer [39].The MBROLA engine is not language specific hence it can be used to develop multilingual outputs. For example MBROLA is used to develop a German TTS since it produces less distortion during signal processing [40].

8.2 FESTIVAL

FESTIVAL provides a general framework for developing speech synthesis systems. It offers full Text to Speech facilities through multiple API's and uses the Edinburgh Speech tool library [41]. FESTIVAL is multilingual and is highly flexible having a modular architecture. FESTIVAL includes modules for text processing, linguistic/prosodic processing and waveform generation. The framework supports di-phone synthesis and unit selection synthesis. It uses a clustering technique for organizing the units in the speech database based on their prosodic and phonetic context [5].

8.3 FLITE

FLITE is a small, fast run-time synthesis library which is suitable for embedded systems. It has been designed as an alternative run-time synthesis platform for Festival where speed of execution and reduced size of TTS system is required. The Library of Flite is much faster and smaller than the similar Festival system [42]. Flite Consists of four modules namely core library module, language module, voice module and language lexicon module [5].

8.4 CHATR

It is a generic speech synthesis engine which is modular in nature and supports unit selection synthesis. Some of the features included in CHATR are multiple types of inputs, support for multiple languages, automatic selection of appropriate units, and choice of waveform synthesis methods, parameterized intonation features, text-to-speech module, and abstract phoneme set to name a few [43]. The speech inventory in CHATR can be viewed as a state transition network, with each speech unit being represented by a separate state [5].

9. CONCLUSION

This paper gives a review about various Concatenative speech synthesis methods. The availability of standard databases, Speech synthesis frameworks and the comparison of different concatenative methods are discussed. The most preferred method of concatenative speech synthesis is the unit selection synthesis, due to naturalness of the synthesized speech. The paper reviews the evaluation techniques for determining the quality of synthesized speech. Although there is a lot of work done, in the field of Concatenative speech synthesis in the past decades, there is scope for further improving the naturalness of the synthesized speech while trying to reduce the footprint of the speech inventory.

10. REFERENCES

[1] Sak, Hasim, Tunga GUNGOR, and Yasar SAFKAN. "A corpus-based concatenative speech synthesis system for Turkish." Turk J Elec Engin 14.2 (2006).

[2] Newton, P.S.R. :Review of methods of Speech Synthesis, EE Dept., IIT Bombay(November 2011)

[3] Tabet Y, and Mohamed Boughazi. "Speech synthesis techniques. A survey." Systems, Signal Processing and their Applications (WOSSPA), 2011 7th International Workshop on. IEEE, 2011.".

[4] Indumathi A., and E. Chandra. "Survey on speech synthesis." Int J Signal Process 6 (2012): 140-5.

[5] Samuel Thomas, "Natural sounding Text-to-speech synthesis based on syllable-like units," M S Thesis, IIT, Madras,2007.

[6] Indumathi, A., and E. Chandra. "Survey on speech synthesis." Int J Signal Process 6 (2012): 140-5.

[7] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan, and Hema A. Murthy, Text-to-Speech Synthesis using syllable-like units Proceedings of National Conference on Communications, IIT, India. 2005.

[8] Narendra, N. P., and K. Sreenivasa Rao. "Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis." Applied Soft Computing 13.2 (2013): 773-781.

[9] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 1, pp. 373–376, 1996.

[10] Heather Cryer and Sarah Home (2010),"Review of methods for evaluating synthetic speech", RNIB Centre for Accessible Information (CAI) Technical report No. 8, 1-12.

[11] Aimilios Chalamandaris, Sotiris Karabetsos, Pirros Tsiakoulis,and Spyros Raptis(2010)," A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers ", IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, 1890-1897.

[12] E.veera raghavendra, srinivas Desai, B.yegnanarayana , Alan W.Black, Kishore Prahallad ,"Global Syllable Set For Building Speech Synthesis In Indian Languages ",in Proceedings of IEEE 2008 workshop on Spoken Language Technologies, Goa, India, December 2008.

[13] John Kominek, Tanja Schultz, and Alan W. Black. "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." In SLTU, pp. 63-68. 2008.

[14] Shinnosuke Takamichi et. al, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis", IEEE Journal Of Selected Topics In Signal Processing, Vol. 8,No.2, April 2014 pp 239-250.

[15] Stas Tiomkin, David Malah, Slava Shechtman, and Zvi Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units" IEEE Transactions on Audio, SPEECH, and Language Processing, vol. 19, no. 5, JULY 2011 pp 1278-1288.

[16] Mandal, Shyamal Kumar Das and Datta, Asoke kumar, "Epoch Synchronous non-overlap-add (ESNOLA) method based concatenative speech synthesis system for Bangla". ISCA workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007

[17] Toma, Ștefan-Adrian, et al. "A TD-PSOLA based method for speech synthesis and compression." Communications (COMM), 2010 8th International Conference on. IEEE, 2010.

[18] Mattheyses, We-sley, Werner Verhelst, and Piet Verhoeve. "Robust pitch marking for prosodic modification of speech using TD-PSOLA." Proceedings of the IEEE Benelux/DSP Valley Signal Processing Symposium, SPS-DARTS. 2006. pp. 43-46.

[19] Schnell, Norbert, et al. "Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA)." Proceedings of the International Computer Music Conference. 2000, pp.102-108

[20] Mukherjee, Sankar, and Shyamal Kumar Das Mandal. "A Bengali speech synthesizer on Android OS." Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments. Association for Computational Linguistics, 2012, pp. 43–46.

[21] Language technological journal of TDIL : vishvabharat Epoch Synchronous Non-Overlapping Add (ESNOLA) Approach Concatenative Text to Speech Synthesis-A Technical Report.[Online]http://tdil.mit.gov.in/

[22] Pammi, Sathish, Marc Schröder, Marcela Charfuelan, Oytun Türk, and Ingmar Steiner. "Synthesis of listener vocalisations with imposed intonation contours." In SSW, 2010, pp. 240-245.

[23] Rao, K. Sreenivasa, and B. Yegnanarayana. "Prosody modification using instants of significant excitation." Audio, Speech, and Language Processing, IEEE Transactions on vol.14, no. ``3 (2006): 972-980.

[24] Heiga Zen, Keiichi Tokuda, Alan W. Black ,"Statistical parametric speech synthesis", Speech Communication vol.51,no.11,2009,pp. 1039–1064.

[25] Raitio, Tuomo, et al. "HMM-based speech synthesis utilizing glottal inverse filtering." Audio, Speech, and Language Processing, IEEE Transactions on vol.19, no.1, 2011, pp. 153-165.

[26] Yu, Kai, Heiga Zen, François Mairesse, and Steve Young. "Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis." Speech communication, vol.53, no. 6, 2011, pp. 914-923.

[27] Lu, H., Ling, Z. H., Wei, S., Dai, L. R., & Wang, R. H.," Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier". In INTERSPEECH, Vol. 10, 2010, and pp. 162-165).

[28] Fu-Chiang Chou, Chiu-Yu Tseng, and Lin-Shan Lee, "A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese", IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, 2002, pp 481-494.

[29] V. Kamakshi Prasad, T. Nagarajan and Hema A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions", Speech Communications, Elsevier publications, Vol.42, 2004,pp.429-446.

[30] Size of Speech Corpora (As on july 2014) , [Online] http://www.ldcil.org/resourcesSpeechCorp.aspx

[31] John Kominek, Alan W Black," THE CMU ARCTIC SPEECH DATABASES", 5th ISCA Speech Synthesis Workshop – Pittsburgh, 2004, pp 223-224.

[32] [Online],CMU ARCTIC speech synthesis databases, http://festvox.org/cmu arctic/

[33] Online],CMU FAF speech synthesis databases, http://festvox.org/cmu_faf/

[34] [Online],CMU SIN speech synthesis databases, http://festvox.org/cmu_sin/

[35] http://festvox.org/dbs/dbs_kdt.html

[36] Catherine Stevens, Nicole Lees, Julie Vonwiller , Denis Burnham ,” On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference”, Computer Speech and Language Elsevier publications Vol. 19, 2005,pp 129–146.

[37] Azis, Nur Aziza, et al. "Evaluation of text-to-speech synthesizer for Indonesian language using semantically unpredictable sentences test: IndoTTS, eSpeak, and google translate TTS." Advanced Computer Science and Information System (ICACSIS), 2011 International Conference on. IEEE, pp. 237-242.

[38] Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark J. F. Gales, Kate Knill, Sacha Krstulovic, and Javier Latorre, “Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization”, ." Audio, Speech, and Language Processing, IEEE Transactions on 20, no. 6 (2012),pp 1713-1724.

[39] http://tcts.fpms.ac.be/synthesis/mbrola.html

[40] Marc Schröder and Jürgen Trouvain. "The German text-to-speech synthesis system MARY: A tool for research, development and teaching." International Journal of Speech Technology 6, no. 4 (2003),pp- 365-377.

[41] Black Alan. Paul Taylor, Richard Caley, and Rob Clark. "The festival speech synthesis system." University of Edinburgh 1 (2002).

[42] Black. Alan. and Kevin A. Lenzo. "Flite: a small fast run-time synthesis engine." 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis. 2001.

[43] Black Alan and Paul Taylor. "CHATR: a generic speech synthesis system." Proceedings of the 15th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1994.

[44] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender